CHILDREN WITH READING DISABILITIES: DOES DYNAMIC ASSESSMENT HELP IN THE CLASSIFICATION?

H. Lee Swanson and Crystal B. Howard

Abstract. **This study was conducted to determine whether the cognitive performance of reading disabled and poor readers can be separated under dynamic assessment procedures, and whether measures related to dynamic assessment add unique variance, beyond IQ, in predicting reading achievement scores. The sample consisted of 70 children (39 females and 31 males). Within this sample four groups of children were compared: children with reading disabilities (*n*=12), children with math/reading disabilities (*n*=19), poor readers (*n*=14), and skilled readers (*n*=25). Intelligence, reading and math tests, and verbal working memory (WM) measures were administered (presented under static and dynamic testing conditions). Two important findings emerged: (a) hierarchical regression analyses found that a dynamic assessment measure factor score contributed unique variance to predicting reading and mathematics, beyond what is attributed to verbal IQ and initial scores related to WM; and (b) poor readers and skilled readers were more likely to change and maintain their WM score gained under the dynamic testing conditions than children with reading disabilities or children with a combination of math/reading disabilities. Implications for a valid classification of reading disabilities are discussed.**

H. LEE SWANSON, Ph.D., is professor, educational psychology and special education, University of California, Riverside. CRYSTAL B. HOWARD is a doctoral candidate, educational psychology, University of California, Riverside.

Children with reading disabilities (RD) experience information-processing difficulties on specific cognitive tasks (e.g., Stanovich & Siegel, 1994; Swanson & Siegel, 2001; Torgesen, 2002). These processing difficulties are assumed to be intrinsic to the child; that is, they are not due to instructional or environmental factors (e.g., Shaywitz et al., 1999). Further, RD children's processing difficulties are reflected in specific academic domains (e.g., reading) that draw upon those processes (e.g., Swanson & Siegel, 2001; Torgesen, 2002). In addition, it is assumed that these specific processing deficits are unexpected given their overall potential (see Fletcher et al., 2002; Stanovich & Siegel, 1994; for a review of assumptions). Given these assumptions, at least two questions emerge.

First, how should "potential" be measured? The notion of potential has played a critical role in defining learning disabilities (LD) since the inception of the field (e.g., see Bateman, 1992, for review). Typically, differences between IQ and achievement on standardized

tests are viewed as a prototype for representing differences between potential and actual performance (Fletcher, Francis, Rourke, Shaywitz, & Shaywitz, 1992; Shepherd, Smith, & Vojir, 1983). However, a review of the literature suggests that such procedures are invalid for classification purposes (e.g., Fletcher et al., 1992; Hoskyn & Swanson, 2000; Stuebing et al., 2002). For example, the relevance of standardized intelligence measures (e.g., WISC-III) in the diagnostic classification of learning disabilities has been criticized because reading achievement within samples with LD is not predicted by variations (high vs. low) in IQ (e.g., Fletcher et al., 2002; Hoskyn & Swanson, 2000; Stanovich & Siegel, 1994; Siegel, 1989, 1992; Stuebing et al., 2002). Further, several authors (e.g., Brown & Ferrara, 1999; Campione, 1989; Embretson, 1992) have suggested that traditional intelligence tests (i.e., tests that measure unassisted performance on global measures of academic aptitude) provide a poor estimate of general ability. These authors argue that because static or traditional approaches to assessment typically provide little feedback or practice prior to testing, failure often reflects children's misunderstanding of instructions more that their ability to perform the task. Thus, whether "potential" is adequately captured on traditional IQ measures presents a conceptual problem.

One possible alternative or supplement to traditional assessment is to measure a child's gain in performance when given examiner assistance. Thus, "potential" for learning new information (or accessing previously presented information) is measured in terms of the distance, difference between, and/or change from unassisted performance to a performance level with assistance. Procedures that attempt to modify performance via examiner assistance in an effort to understand learning potential are called dynamic assessment (e.g., see Grigorenko & Sternberg, 1998; Swanson & Lussier, 2001). Although *dynamic assessment* is a term used to characterize several distinct approaches (see Grigorenko & Sternberg, 1998; Swanson & Lussier, 2001; for a review) it includes two critical features: to determine the learner's potential for change when given assistance, and to provide a prospective measure of performance change independent of assistance (Embretson, 1987). Unlike traditional testing procedures, score changes due to examiner intervention are not viewed as threatening task validity. In fact, some authors argue that they increase construct validity (e.g., Carlson & Wiedl, 1979; Elliot & Lauchlan, 1997; Swanson, 1992).

Although dynamic assessment has been suggested as an alternative to traditional assessment (e.g., Day, Engelhardt, Maxwell, & Bolig, 1997; Jitendra & Kameenui, 1993), there are no published data, to the authors' knowledge, on whether children with RD are more sensitive than other ability groups to such procedures. Thus, a number of questions need to be addressed if such procedures are to be used to assess RD. For example, can children with RD, when given instructional support on processing tasks, be differentiated in performance from poor and average readers? This question is important because the processing difficulties of children with RD are assumed to be stable compared to other processing abilities (see Swanson & Hoskyn, 1998, for discussion). Thus, if the processing performance of children with RD can be substantially modified and their performance is statistically comparable to that of normally achieving children, the "intrinsic nature" of RD needs to be reexamined.

Another question is whether children with RD can be separated from poor readers. This issue is important because assessment practices that rely heavily on psychometric tests for classification of children with RD have not provided, to date, systematic procedures for separating those children who primarily have reading problems related to inadequate or weak instructional support from children who have information processing deficits (Torgesen, 2002). Related to this issue is the finding that the cognitive profile of children with RD cannot always be discriminated from that of generally low-achieving children when using static or traditional assessment (Hoskyn & Swanson, 2000; Stuebing et al., 2002).

In summary, the present study had two purposes. First, a determination was made as to whether processing "potential" via dynamic assessment is related to reading achievement. Processing potential is defined as the score obtained with examiner assistance (i.e., gain score) and sustained performance without assistance (i.e., maintenance score). In statistical terms, the question is whether gain and maintenance scores contribute unique variance to reading achievement beyond what is contributed by a traditional intelligence measure. Linking dynamic assessment with reading achievement as well as determining whether "potential" as measured on a commonly used IQ test differs from potential as measured under dynamic testing conditions in the prediction of achievement are important issues if dynamic assessment is to be taken seriously as a valid assessment procedure in the diagnosis of RD.

The tasks used in this study for assessing information processing potential were related to working memory (WM), a critical component of major information-processing models (e.g., Baddeley & Logie, 1999) that has been found to be seriously deficient in children with RD (e.g., De Beni, Palladino, Pazzaglia, & Cornoldi, 1998; Siegel & Ryan, 1989; Swanson, 1993, 2003; also Swanson & Siegel, 2001, for a comprehensive review).[1] All major information-processing models involving skill

acquisition and learning include the component of WM (e.g., see Daneman & Merikle, 1996, for a review), because it is highly correlated with performance on several academic and language-related tasks, such as vocabulary (e.g., Baddeley, Gathercole, & Papagno, 1998), reading comprehension (e.g., Swanson, 1999), language acquisition (e.g., Baddeley et al., 1998), problem solving (e.g., Kyllonen & Christal, 1990), mathematics (e.g., Bull, Johnston, & Roy, 1999), fluid intelligence (e.g., Engle, Kane, & Tuholski, 1999), and writing (McCutchen, 2000). Correlations between WM and reading or intelligence measures with adult samples are in the range of .55 to .92 (e.g., see Daneman & Merikle, 1996).

The standardized test (*N*=1594) used to measure WM was the Swanson-Cognitive Processing Test (S-CPT; Swanson, 1995a). As indicated by Grigorenko and Sternberg (1998), this is one of the few tests that report validity and reliability data. It is an individually administered battery that is assumed to measure different aspects of WM ability and processing potential. Working memory is defined in this test as concurrent processing and storage activities, whereas potential, via dynamic assessment, is defined as (a) learner performance change relative to initial performance on WM measures when given assistance (gain) and (b) performance change independent of assistance (maintenance).

Second, it was of interest to determine whether children with RD can be discriminated via dynamic assessment from children who are poor readers. This is important because several studies (see synthesis of the literature by Hoskyn & Swanson, 2000; Stuebing et al., 2002) indicate that there are no clear psychometric and processing distinctions between poor readers and children with RD. However, the fact that current practices using static measures do not distinguish children with RD from children who are poor readers does not mean it cannot be done. Thus, we examine whether a child's response to assisted performance provides a frame of reference for separating children who are poor readers from children who are RD.[2] Although not related to dynamic assessment, a comprehensive synthesis of the treatment intervention literature indicated that the magnitude of treatment outcomes (effect size) for children with RD was smaller (i.e., they were less responsive) than for poor readers (see Swanson & Hoskyn, 1998, p. 307, for discussion). Based on these findings, it is possible that poor readers will be more responsive to measures of change than children with RD.

In summary, the purpose of the present study was twofold: (a) to determine whether dynamic assessment adds unique variance beyond IQ in predicting reading achievement scores; and, (b) to compare children classified as RD with skilled and poor readers on dynamic assessment measures. It was hypothesized that (a) dynamic assessment measures will contribute significant variance in predicting reading and (b) children with RD will be less responsive to dynamic assessment than children who are poor readers.

METHOD

Participants

The sample consisted of 70 children (39 females and 31 males), primarily drawn from children tested in Southern California. Initial sampling included children with reading difficulties currently receiving special education services in either a public or private school. Children in these settings had been classified as learning disabled (LD) according to state guidelines that closely matched the *Federal Register* definition (1977). Specifically, the definition reflected the following: (a) the learning problem was specific, generally confined to one or two academic areas; (b) the child's poor achievement was not commensurate with his/her ability as in other academic areas which are average or above based on the child's chronological age; and (c) the learning difficulty was not primarily the result of retardation, poor teaching, or cultural deprivation.

From this pool of participants (*N*=203), further selection included identifying children operationally classified by the researchers as reading disabled, reading disabled and math disabled, or poor readers. Our classification of RD followed the "cut-off" scores detailed by Fletcher (Fletcher et al., 1992, 1994) and Siegel (1989; Siegel & Ryan, 1989). All children were administered the reading and math subtests from the Wide Range Achievement Test-Revised (WRAT-R; Jastak & Wilkinson, 1984).[3] The WRAT-R was administered rather than the WRAT-III because the majority of studies that have used cut-off scores to discriminate between poor readers and children with RD have relied on the former measure (e.g., Siegel, 1992; see Hoskyn & Swanson, 2000, for a review). Intelligence scores were measured on the Wechsler Intelligence Scale for Children, Third Edition (WISC-III; Wechsler, 1991).

Operational criteria for RD included a Verbal Scale IQ score > 85 and a word recognition score on the WRAT-R below the 25th percentile (standard score of 90). As expected, children varied tremendously on math skills. Thus, RD children with math scores below a standard score of 86 were considered to have both math and reading disabilities (MD/RD), whereas those with math scores above an 85 standard score were considered the RD group. Verbal IQ was selected as a classification measure because a comprehensive meta-analysis comparing RD and poor readers found that these scores (in contrast to full-scale scores) moderated differences between the two groups (Hoskyn & Swanson, 2000). A

cut-off score in reading rather than a 15-point discrepancy score between reading and IQ was selected because the latter scores have been found to have weak discriminant validity in separating the cognitive performance of children with RD from that of poor readers (Fletcher et al., 1992; Hoskyn & Swanson, 2000; Stuebing et al., 2002).

Children classified as poor readers (*n*=14) followed the operational definition outlined by Hoskyn and Swanson (2000, see p. 105). Criteria for poor readers were defined as children with verbal intelligence scores below a standard score of 96, but above 70, and word recognition and arithmetic scores below the 40th percentile.[4] Skilled readers were sampled from the same school as the less skilled readers. Criteria for skilled readers were reading and math scores above a standard score of 90 on the WRAT-R. Children classified as skilled readers had reading, math, and Verbal scale IQ scores above a standard score of 90.

The final selection (*N*=70) included children from low- (35%) and middle- (65%) income schools. The ethnic re-presentation of the participants was as follows: European American descent: 58; African American: 12; and Asian American: 1. All children were monolingual in English, and less skilled readers had received special education resource room services for at least one year.

Within the final sample of 70, four groups of children were compared: skilled readers (*n*=30), poor readers (*n*=14), reading disabled (RD) (*n*=12), and children with both math and reading disabilities (*n*=19). The mean intelligence, reading, and math scores, and chronological age for each group are shown in Table 1. No significant differences emerged between ability groups in terms of

Table 1
Classification and Performance Measures as a Function of Ability Groups

	Ability Level									
	Total Readers (*n*=70)		**Poor Readers** (*n*=14)		**Skilled Readers** (*n*=25)		**RD** (*n*=12)		**MD/RD** (*n*=19)	
Variables	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
Intelligence	108.02	14.42	89.71	3.29	118.00	11.67	107.23	11.55	107.68	10.05
Reading	86.19	28.20	68.03	13.31	115.76	13.73	70.88	14.69	70.33	26.07
Math	91.51	20.62	76.72	14.08	107.00	13.74	104.90	16.64	73.43	10.05
Age	11.93	2.29	12.26	2.49	12.79	3.84	10.37	1.84	11.53	1.86
Initial										
Rhyming	2.25	1.09	2.00	.96	2.40	1.19	1.83	1.11	2.52	1.02
Digit/Sentence	2.30	1.38	1.92	.91	3.80	1.32	1.50	.90	2.42	1.74
Gain										
Rhyming	3.21	1.08	2.78	.69	3.64	1.46	3.08	.51	3.05	.84
Digit/Sentence	3.90	1.62	3.00	1.03	4.68	1.81	3.83	1.19	3.57	1.57
Maintenance										
Rhyming	2.70	1.35	2.07	1.14	3.28	1.56	2.33	1.15	2.63	1.07
Digit/Sentence	2.84	1.54	2.28	.99	3.70	1.57	2.00	1.21	2.62	1.57
Probe										
Rhyming	3.57	2.30	3.42	1.60	3.72	2.40	3.83	2.79	3.50	2.35
Digit/Sentence	2.89	3.05	2.21	2.45	3.92	3.27	4.66	3.20	.83	1.50

gender, χ^2 (3, N=70) = 3.07, $p > .05$; ethnicity, χ^2 (9, N=70) = 12.58, $p > .05$; or chronological age, F (3,66) = 2.14, $p > 05$. A Tukey test indicated no significant differences ($ps > .05$) in reading scores between poor readers, children with RD or the MD/RD group. However, a significant advantage ($p < .05$) was found for the RD group when compared to the MD/RD group on math scores, and when compared to the poor readers on verbal IQ.

Measures

Two subtests from a battery of 11 of the S-CPT were selected because they represent verbal WM processing. One is assumed to tap components of phonological WM (rhyming task) and the other to tap components of semantic (digit/sentence task) memory. The WM subtests conformed to Baddeley's (1986) definition that they "require simultaneous processing and storage of information" and "measure various contents" (pp. 34-35). A critical feature of the WM tasks was that they required the maintenance of some information during the processing of other information. Consistent with Daneman and Carpenter's WM measures (1980), the processing of information was assessed by asking participants simple questions about the to-be-remembered material (storage + processing demands), whereas storage was assessed by accuracy of item retrieval (storage demands only). The process question generally requires a simple recognition of new and old information, and is analogous to Daneman and Carpenter's task, which requires a yes/no response to previously presented information. It is important to note, however, that the difficulty of the processing question remained constant within task conditions, thereby allowing the source of performance differences to reflect increases in storage demands. The Cronbach alpha for each task used in this study, with age partialed out, was > .80 (Swanson, 1995a, 1995b).

Verbal WM

Rhyming. The purpose of this task was to assess the child's recall of acoustically similar words. The child listened to sets of words that rhyme. Each successive word in the set was presented every two seconds. There were nine word sets, ranging from 2 to 14 monosyllabic words. The dependent measure was the number of sets recalled. Before recalling the words, the child was asked whether a particular word was included in the set. For example, the child was presented the words "lip-slip-clip," and then asked if "ship or lip" was presented in the word set. The child was then asked to recall the previously presented words (lip-slip-clip) in order. The dependent measure was the number of sets recalled correctly (range of 0 to 9).

If the child omitted, inserted, or incorrectly ordered the words, a series of probe responses was presented, which continued until the child could no longer provide the correct response. For example, for the sample item "car-star-bar-far" (Item #3) and the process question "Which word did I say – jar or star?" consider the probe sequence:

1. The last word in the sequence was "far," now can you tell me all the words in order?
2. The first word in the sequence was "car," now can you tell me all the words in order?
3. The middle words in the sequence are "star" and "bar," now can you tell me all the words in order?
4. All the words in order are "car-star-bar-far," now can you tell me all the words?

For each set of items not recalled in the correct order or for items left out or substituted, the experimenter provided a series of hints based on the error that was closest to Probe 1. That is, probes went from the least obvious hint (Probe 1) to the next explicit hint that facilitated recall of the answer. Once the appropriate hint had been identified based on the location of the error, probes were presented in order until the correct sequence was given. For example, suppose the child for Item #3 responded car-bar-far. The child obviously left out a word in the middle, so the experimenter would provide a hint related to the middle words (Probe 3, in this case). If Probe 3 did not provide the correct response, the experimenter moved to Probe 4. In contrast, if a child responded initially by saying only car, the sequence began with Probe 1 and proceeded through all probes until the correct response was given. If a correct response did not occur after probing, the task was discontinued, and the next task was administered. If a correct response did occur, the next set of items of increased difficulty was presented.

Digit/sentence task. The purpose of this task was to assess the participant's ability to remember numerical information embedded in a short sentence. The administration of items and probes followed the same format as the rhyming task. Prior to stimulus presentation, the participant was shown a figure (see Swanson, 1993, Figure 1) depicting four strategies for recalling numerical information. These strategies were pictorial representations of rehearsal, chunking, associating, and elaborating of information. The general instructions for introducing the strategies were as follows:

> "I'm going to read you some sentences that have information I want you to remember. All the sentences have to do with remembering an address, but I would like you to pay attention to all the information in the sentence because I will ask you a question about the sentence. After I present this information, and before you recall it, I will ask you to choose a strategy (for children under ten – the phrase, "A way

of remembering the information" was used) that you think will best help you remember."

The experimenter then showed four pictures, each depicting a person thinking about using one of the four strategies (see Swanson, 1993). As the experimenter explained each strategy, he/she pointed to the picture that matches the description. The experimenter stated that

> "Some of the ways that may help you remember are: (1) saying the numbers over to yourself. For example, if I say '2-4-6-3 Bader Street,' you would say to yourself '2-4-6-3' over and over again, or (2) you might say some numbers together in pairs. For example, if I say the numbers '2-4-6-3 Bader Street,' you would say '24 and 63,' or (3) you may just want to remember that the numbers go with a particular street and location. For example, if I say '2-4-6-3 Bader Street,' you would remember that 2-4-6-3 and Bader Street goes together, or (4) you might think of other things that go with the numbers. For example, if I say '2-4-6-3' you might think 2-4-6-3 I have to go climb a tree."

These four pictorial representations of strategies generally reflect rehearsal, chunking, associating, and elaborating of information, respectively. After all strategies had been explained, participants were presented with item sets that included numbers in a sentence context. They were then told that they must recall the numbers in the sentence in order shortly after they select from (point to) a pictorial array representing the strategy that best approximates how they will attempt to remember the information. No further information about the strategies shown in the picture was provided. Participants were allowed 10 seconds to make a decision. The range of recall difficulty was 3 digits to 14 digits, and the dependent measure was the highest number of sets correctly recalled (range of difficulty 0 to 9).

Thus, the sequence of the steps for administration after the introduction was as follows: (a) the participants were orally read a sentence (the numbers in the sentence were presented at the rate of approximately one every two seconds); (b) participants were asked a process question that required them to give the name of the street referred to in the target sentence; (c) participants were asked to select one of the four strategies that were represented pictorially that were most like the one they would use to remember the order of the street numbers; (d) participants were asked to recall the numbers of the address in the order in which they were originally presented; and (e) if an error in recall occurred, the probe questions were implemented.

Probing procedures followed the same format as the rhyming task: hints were provided sequentially based on the type of error, ranging from least obvious hint (Probe 1) to the next explicit hint that facilitated recall of the answer. If probing did not elicit a correct response, the task was discontinued and the next task was administered. If a correct response did occur, the next set of items of increased difficulty was presented.

Reliability and Validity of Measures

The tasks above were selected because of their high reliability and validity. For example, the tasks correlated significantly with the Sentence-Span task (Swanson, 1996), a seminal measure of WM (Daneman & Carpenter, 1980). Further, previous studies using an independent sample had shown that the rhyming and the digit/sentence task correlated significantly with reading comprehension, word recognition, mathematics and word problem solving (Swanson, 1995a, 1996). Reliability of the tasks ranged from .76 to .85 with the current sample. The reliability (coefficient alpha) for the sample was .85 for initial, .81 for gain, and .76 for maintenance for the rhyming task, and .83 for initial, .79 for gain, and .78 for the digit/sentence task.

The tasks were also selected because they reflected the accessing of different information from LTM. The rhyming task taps sequential order for phonological information, whereas the digit/sentence task taps the reorganization of words into categories. Both tasks, however, have a similar format of introducing the items to be remembered, followed by a process question, and finally a storage question. In addition, both tasks reflected the introduction of increasingly difficult sets of information to be remembered. The digit sentence task also asked participants to select a picture prior to retrieval. It was decided to keep this part of the task in order to follow the standardized instructions.

Achievement Measures and Verbal IQ

The Reading and Mathematics subtests from the WRAT-R (Jastak & Wilkinson, 1984) were used as the criterion measures. The reading subtest contains tasks of naming single words, and the arithmetic subtest involves solving written computations. Median reliability across groups for each subtest was .92. All children were individually administered subtests of the verbal section of the WISC-III (Wechsler, 1991). The WISC-III measures contain 13 subtests of which 3 are supplementary. The standard verbal subtests consist of Information, Similarities, Arithmetic, Vocabulary, and Comprehension.

Dependent Measures

There were two primary dependent measures. One score reflected the recall of increasingly difficult sets of items under three conditions: initial, gain, and maintenance. The most difficult set recalled in each condition was referred to as an initial, gain or maintenance span score. The second measure was a probe score. This

referred to the number of hints provided. We briefly describe below the rationale for these measures.

The measures were selected to address the issue of which type of scores most accurately measure processing potential and predict reading. Several authors consider the first area of focus in assessment to be one of improving the processing of information. For example, utilizing Vygotsky's (1978) zone of proximal development, Brown and French (1979) made a distinction between an individual's proximal potential and actual level of performance. In the area of child development, for example, they state:

> A distinction is made between a child's actual development, i.e., his/her completed development as might be measured on a standardized test, and his/her level of potential development, the degree of competence he/she can achieve with aid. Both measures are seen as essential to diagnosis of learning abilities and the concomitant design of remedial programs. (p. 210)

In the S-CPT, the "zone of potential" was assessed by determining optimal memory performance. This consisted of determining the number of probes or hints necessary to enhance the examinee's access to previously stored information. An assessment of the examinee's potential (i.e., ability to access available information) involved three steps. First, the examinee was administered a battery of items on a particular subtest. Second, if the examinee failed to retrieve the item information, the examiner provided a series of progressive probes based upon the information that was forgotten. The number of probes or hints (probes) necessary to achieve maximal performance was considered "the width" of the individual's zone of potential. Third, the items at which the examinee achieved the highest level of performance were readministered at a later time. This "maintenance" activity was important because it reflected the examinee's ability to benefit from the "aids" or probes provided by the examiner. The examinee's ability to maintain behavior provides valuable assessment information about the potency of the aids that help the examinee access information.

As stated, a major goal of dynamic assessment models is to show not that one can better estimate ability, but to measure modifiability (Embretson, 1987; Grigorenko & Sternberg, 1998; Swanson & Lussier, 2001). A major issue here is the type of scores necessary to measure modifiability (see Embretson, 1987, for a review). For example, Campione and Brown (1987) measured modifiability as the number of hints needed to solve a problem that has been failed. Thus, the fewer the hints, the more modifiability the examinee possesses. Embretson (1987) has suggested that this score merely provides a better estimate of initial ability (see p. 149). Another method to measure modifiability was to bring scores to an asymptotic level (under the probing conditions) and then obtain a measure on the subtest again after the probes had been removed. The basic rationale was to eliminate performance differences due to different strategies or unfamiliarity with the laboratory procedures. As yet, there is no agreed-upon measure of cognitive modifiability (Grigorenko & Sternberg, 1998; Swanson & Lussier, 2001). To partially address this issue, several measures were used to determine which measure best predicts WM change.

The first DA measure, Gain Score or asymptotic level, was the highest score obtainable under probing conditions. A second measure, Maintenance Score, was the stability of the asymptotic level after the probing conditions had been removed. This measure was scored dichotomously in that the Gain Score was either maintained or not. In cases were a Gain Score was not maintained, the Initial Score (initial performance) was assigned to the examinee. Thus, modifiability was measured in an absolute sense. A third scoring procedure was the number of hints needed to achieve the Gain Score. Thus, a Probe Score was the number of prompts necessary to achieve the asymptotic level.

In establishing the Gain and Maintenance scores after the Initial Score, probes that matched the corresponding error were administered. After beginning with the appropriate probe, probes were administered in sequential order until the correct response was achieved. The examiner recorded the number of probes that led to the Gain Score. If the Initial Score equaled the gain score, the Probe Score was zero. If all four probes in an item set, or three probes in two consecutive item sets, had been administered, the examiner moved to the next subtest.

Procedure

Individual testing was performed by either interns in a school psychology program or graduate students in test and measurement classes. Students were familiar with and had previously administered all traditional measures. Each examiner received a special three-hour single-session training unit for the dynamic measurement prior to testing the children. The student testing was completed in one session with a total test time of approximately 40 minutes per child. All subtests were administered following the instructions in the standardization.

All items for the initial condition were administered until (a) a process question was missed or (b) an error in retrieval occurred. If an error in retrieval occurred (a participant omitted, inserted, or incorrectly ordered the numbers, dots, words, related to the appropriate task), probes were administered. The only stipulation for insti-

tuting the probing condition was that the process question be answered correctly. Probes were administered based on the type of error made (i.e., whether the error was related to recency, primacy, or middle items), and probing procedures continued until all targeted items could be recalled correctly. The "bow-shaped curve," commonly found in episodic memory studies (i.e., items presented at the end of the list are remembered better at those at the beginning and those items presented at the beginning of a list are better remembered than those presented in the middle), provides the basis for ordering a series of cues from implicit to explicit information (see Swanson, 1995a, pp. 5 -9, for rationale on ordering of probes). Cues (the terms *cues* and *probes* are used interchangeably) were administered based on the type of error made (i.e., whether the error was related to recency, primacy, or middle items), and cuing procedures continued until none of the targeted items could be recalled.

The order of cues was based on the assumption that the first cue provides information about the final items because these items are the least susceptible to interference. The second cue was assumed to provide information about the primacy (first) items because they are the most reliant on long-term memory processes. The third cue provided additional information about the middle-presented items because these items are the most susceptible to interference and storage limitations. Finally, if the participant failed to benefit from any of the previous three cues, all the items were repeated and retested. Probing procedures continued until none of the targeted items could be accessed (recalled). Because participants were only probed about items for which they answered the process question correctly, it was assumed that poor item retrieval was attributable to item accessibility rather than to items not being adequately stored.

After the two subtests from the S-CPT were administered under initial and gain conditions, participants were readminstered the items for the highest successful set (highest set of items established under gain conditions) for each task. The general instructions were "These items were presented to you earlier. I want to see what you can remember this time without hints."

Order effects were tested in an earlier pilot study. An analysis of variance on the factor of presentation order was not significant, $F < 1$. Thus, the rhyming tasks were administered first followed by the digit-sentence span.

RESULTS

Table 1 shows the means and standard deviations of scores for verbal IQ, reading, and the cognitive measures as a function of each subtest and testing condition across the four ability groups.

Group Comparison

The means and standard deviations for raw span scores of the four reading ability groups are shown in Table 1. In addition, the effect sizes (η^2) related to group differences are also reported. An η^2 of .13 and .05 is equivalent to a Cohen's d of .80 and .50, respectively. An alpha of .05 was utilized for all comparisons presented below unless stated otherwise. The critical overall *F*-value for achieving a power of .80 at alpha = .05 was 3.80 (see Murphy & Myors, 1998, Appendix B, for $df = 3, 66$). A MANOVA collapsing performance across conditions and tasks indicated that the sample size was adequate to detect ability group differences, $F(3,66) = 4.63$, $p < .01$, $MSE = .501$, $\eta^2 = .17$. We now consider each task separately because of differences in scaling.

Rhyming. A 4 (group) x 3 (condition: initial, gain, maintenance) was computed on span scores, with repeated measures on the last factor. No significant main effect emerged for ability group, $F(3,66) = 2.42$, $p > .05$, $MSE = 3.09$, $\eta^2 = .10$. However a significant effect emerged for condition, $F(3,134) = 32.90$, $p < .001$, $MSE = .45$, and the ability group x condition interaction, $F(2,134) = 2.21$, $p < .05$, $MSE = .45$. A test of simple effects indicated that group differences emerged on the maintenance, $F(3,66) = 3.10$, $p < .05$, $MSE = 1.68$, $\eta^2 = .12$, but not the initial, $F(3,66) = 1.39$, $p > .05$, $MSE = 1.18$, $\eta^2 = .06$, or gain condition, $F(3,66) = 2.30$, $p > .05$, $MSE = 1.68$, $\eta^2 = .10$. Tukey tests indicated that skilled readers outperformed the other groups (skilled readers > poor readers = RD = MD/RD) on the maintenance condition. That is, although the skilled readers outperformed the other groups on the maintenance condition, the performance of poor readers, children with RD, and children with RD and MD was statistically comparable. A Tukey test also showed that significant differences ($ps < .05$) in patterns emerged between conditions for skilled readers (Gain = Maintenance > Initial), as well as poor readers, and children with RD and MD/RD (Gain = Maintenance > Initial).

Digit/sentence. A 4 (group) x 3 (condition: initial, gain, maintenance) was computed on span scores, with repeated measures on the last factor. A significant main effect emerged for ability group, $F(3,66) = 4.85$, $p < .01$, $MSE = 4.61$, $\eta^2 = .18$, and condition, $F(3,134) = 59.18$, $p < .001$, $MSE = 2.36$. The ability group x condition interaction was also significant, $F(2,134) = 2.36$, $p < .05$, $MSE = 2.36$. A test of simple effects indicated that group differences emerged on the initial, $F(3,66) = 3.05$, $p < .05$, $MSE = 1.76$, $\eta^2 = .12$, gain, $F(3,66) = 4.09$, $p > .05$, $MSE = 2.32$, $\eta^2 = .16$, and maintenance condition, $F(3,66) = 5.48$, $p < .01$, $MSE = 2.00$, $\eta^2 = .20$. Tukey tests indicated that skilled readers outperformed the other groups on initial (skilled readers > MD/RD > RD = poor readers), gain (skilled readers > MD/RD = RD > poor

readers) and maintenance conditions (skilled readers > MD/RD = RD = poor readers). A Tukey test showed that significant differences (*ps* < .05) in patterns emerged between conditions for skilled readers (Gain = Maintenance > Initial), poor readers and MD/RD children (Gain > Maintenance = Initial), and children with RD (gain > maintenance > initial).

Probe scores. A MANOVA indicated that significant differences emerged for probe scores across the rhyming and digit/sentence task, $\Lambda = .76$, $F(6,128) = 3.10$, $p < .01$. A significant ANOVA emerged for ability group on the digit/sentence, $F(2,65) = 6.53$, $p < .0001$, $MSE = 7.49$, $\eta^2 = .23$., but not the rhyming task, $F(2,65) = .10$, $p > .05$, MSE = 5.42, $\eta^2 = .01$. A Tukey test showed that the MD/RD group was significantly less responsive to probes than skilled and RD children (MD/RD < skilled reader = RD; skilled reader = poor reader = RD).

To determine if probe procedures influenced performance within groups, effect sizes were calculated between the gain and initial condition. Mean effect sizes ([gain span score-initial span score]/SD of initial score) were calculated across the two tasks to determine the magnitude of change. Cohen (1988) considered effect sizes that approximate .80 as substantial, those that approximate .60 as moderate, and those that approximate .40 as weak. The mean effect sizes for poor readers, skilled readers, RD, and MD/RD children for the rhyming task were .82, 1.04, 1.12, and .51, respectively. The mean effect sizes for poor readers, skilled readers, RD, and MD/RD for the sentence/digit task were 1.16, 1.42, 2.57, and .66, respectively. Thus, only the MD/RD group failed to improve substantially across gain condition for both tasks.

A different picture emerged when effect sizes were calculated between initial and maintenance conditions ([maintenance span score-initial span score]/SD of initial score). The mean effect sizes for poor readers, skilled readers, RD, and MD/RD for the rhyming task were .07, .73, .44 , and .10, respectively. The mean effect sizes for poor readers, skilled readers, RD, and MD/RD for the sentence/digit task were .38, .70, .55, and .12, respectively. Thus, only the skilled readers were able to approximate a .80 effect size on the maintenance condition. (As one reviewer indicated, however, this effect size [.80] was inflated because there was no correction for the autocorrelation. The corrected effect size was .51.) In addition, the magnitude of the effect sizes within groups was higher for children with RD than poor readers and children with MD/RD.

Strategy Selection

Because ability group differences emerged on the digit/sentence task, it may be argued that the advantage of some children was due to a metacognitive knowledge of strategies. A chi-square analysis was computed on strategy choice as a function of group. The results indicated no significant differences on the digit/sentence task, χ^2 (12, $n = 70$) = 12.76, $p > .05$. All four groups were more likely to select pictorial representations of rehearsal or clustering that the other choices. Taken together, the findings on strategy choice do not appear related to performance. This is because group differences in span scores emerged on the digit/sentence task, but strategy choices did not differ between the groups.

Change Analysis

Because performance stability and change between groups were a major focus of the study, the results were analyzed to identify which participants were able to attain a maintenance score higher than their initial scores. That is, although the above analysis provides

Table 2

Percentage of Participants Who Demonstrated Change from Initial to Maintenance Condition

	No Change Score	Change Rhyming Task	Change Digit/Sentence	Change Both Tasks
Poor Readers	21.43	35.71	50.00	7.14
Skilled Readers	36.67	60.00	48.00	36.00
RD	57.69	33.33	25.00	8.33
MD/RD	68.42	5.26	31.58	5.26

Table 3
Intercorrelation Matrix (N=70)

		1	2	3	4	5	6	7	8	9	10	11
1. Initial Rhyming		—										
2. Initial Dig/Sent.		.42	—									
3. Gain Rhyming		.56		.44	—							
4. Gain Dig/Sent.		.40	.64	.44	—							
5. Maintenance Rhyming		.63	.34	.73	.42	—						
6. Maintenance Dig/Sent.		.41	.65	.37	.70	.38	—					
7. Probe Rhyming		.07	.02	.51	.03	.17	-.14	—				
8. Probe Dig/Sent.		-.37	-.15	-.05	.34	-.01	.01	-.02	—			
9. Verbal IQ Score		.06	.29	.36	.30	.31	.33	.06	.14	—		
10. Reading Score		.02	.34	.30	.43	.24	.45	-.02	.33	.56	—	
11. Math		-.03	.16	.38	.42	.38	.30	.04	.35	.42	.56	—
12. Age		.38	.37	.45	.41	.40	.30	-.07	.12	.17	.16	—
Skewedness	.02	.16	.35	-.24	.35	.23	.35	.15	.41	.13	.25	.73
Kurtosis	-.59	-.01	1.54	-.19	-.29	.13	-.37	-.46	-.69	-.64	-.21	.83

If $r > .38$, then $p < .001$.

information on the magnitude of change, it does not include information on the percentage of children who changed their performance from the initial to the maintenance condition. For this analysis, participants whose maintenance scores were higher than their initial score on either the rhyming or sentence/digit task were assigned a score of 1 (change on one task) or 2 (change on both tasks) and those participants whose maintenance scores matched their initial scores for both tasks were assigned a score of 0.

Table 2 lists the distribution of scores. Column 1 shows the percentage of children whose maintenance score was the same as their initial score. In contrast, column 4 shows the percentage of children who maintained a higher score on the maintenance than initial condition across both tasks. The middle columns show the percentage of children with higher scores for the maintenance than the initial condition for the rhyming (column 2) and digit/sentence task (column 3). As illustrated, the ability groups varied in the distribution of change scores, χ^2 (6, N = 70) = 18.37, $p < .01$. Participants who frequently showed the greatest amount of change were poor readers and skilled readers. Approximately 78% of the poor readers and 72% of the skilled readers were more likely to receive a change score of 1 or 2 compared to children with RD (49%) and MD/RD children (31%). When groups were compared separately, MD/RD children were less responsive to

change than poor readers, χ^2 (1, N = 33) = 7.31, p < .01, and skilled readers, χ^2 (1, N = 44) = 7.11, p < .01, respectively. No other significant effects emerged.

Taken together, the results showed that the changes in WM performance of children with RD/MD were distinct from those classified as poor and skilled readers.

In summary, an advantage was found for the skilled readers in WM performance across all conditions compared to poor readers and children with RD or MD/RD. Consistent with the literature, poor readers and children with RD or with MD/RD could not be adequately differentiated on memory scores on either the rhyming or the digit/sentence task. However, participants varied in change scores. Approximately 78% of the poor readers and 72% of the skilled readers were more likely to receive a change score compared to children with RD (49%) and MD/RD children (31%). Two additional findings are important: (a) WM performance for all groups improved on the gain condition, but the smallest magnitude of change emerged for the children with MD/RD; and (b) ability group effects for the rhyming and digit/sentence task were larger for the maintenance (.12, .20) than gain (.10 , .15 and initial conditions (.06, .12), respectively.

Correlation/Hierarchical Regression

The subsequent analysis determined if verbal intelligence, WM, and testing conditions contributed unique variance to reading performance. In addition, to make comparisons between the contributions of these variables, mathematics performance was also considered as a criterion measure. The intercorrelations among verbal IQ, reading, and WM measures as a function of test administration format (initial, gain, maintenance) are shown in Table 3. As illustrated, all measures meet standard criteria for univariate normality (Kline, 1998) with skewedness for all measures less than 3 and kurtosis for all measures less than 4. Also, the data were screened for both univariate and multivariate outliers. Univariate outliers were defined as cases that were more than 3.5 standard deviations from the means. Multivariate outliers were examined by calculating Mahahalobis' d^2. None of the cases were deemed outliers.

For data reduction purposes, as well as control for multicollinearity, factor-score scores were computed

Table 4
Principal-Factor Analysis of Predictor Variables (Varimax Rotation)

	Semantic	Phonological	Semantic-Resp.	Phon-Resp.	Verbal-IQ
Verbal IQ	.25	.12	.13	.08	.47
Initial					
Rhyming	.36	.71	.32	.06	.07
Dig/Sent.	.75	.16	-.17	.07	.19
Gain					
Rhyming	.30	.59	-.03	.56	.31
Dig/Sent.	.78	.29	.35	.05	.07
Maintenance					
Rhyming	.22	.74	.02	.17	.27
Dig/Sent.	.74	.25	.001	-.15	.24
Probe					
Rhyming	-.06	.09	-.01	.71	.03
Dig/Sent.	.02	-.11	.79	-.01	.10

Note. Resp. = Responsive.

Table 5
Hierarchical Regression for Variables Predicting Reading (N=70)

	B	*SE*	*β*	*t*-ratio
Model 1				
Sem. WM	.44	.12	.39	3.46**
Phon. WM	.03	.13	.03	.28
Model 2				
Semantic Responsiveness	.36	.13	.30	2.63*
Phonological Responsiveness	.04	.14	.03	.30
Model 3				
Verbal IQ	.84	.16	.52	5.14***
Model 4				
Semantic WM	.34	.10	.30	3.19***
Phonological WM	.01	.11	.01	.14
Semantic Responsive	.27	.11	.23	2.47*
Phonological Responsive	-.05	.12	- .04	-.47
Verbal IQ	.73	.15	.46	4.74**
Model 5				
Semantic WM	.37	.11	.34	3.40***
Phonological WM	.02	.11	- .01	.14
Verbal IQ	.77	.15	.48	4.85**

Note. Model 1, R^2 = .16, $F(2,67) = 6.35$, $p < .01$; Model 2, R^2 = .10, $F(2,67) = 3.53$, $p < .05$; Model 3, R^2 = .28, $F(1,68) = 26.39$, $p < .001$; Model 4, R^2 =.44, $F(5,64) = 10.05$, $p < .0001$; Model 5, R^2 = .38, $F(3,66) = 13.58$, $p < .0001$.
* $p < .05$, ** $p < .01$, *** $p < .001$.

based a common factor analysis. A common factor analysis was used because it examines only reliable variance among the variables and adjusts for measurement error (see Nunnally & Bernstein, 1994, pp. 522-524, for discussion). Along with the various WM scores, verbal IQ was also included. This was based on recent findings suggesting that WM may be a surrogate measure of intelligence (Engle et al., 1999). Table 3 shows that five factors emerged.

To determine if the five-factor structure was an adequate extraction of the matrix, maximum-likelihood estimates (Joreskog & Sorbom, 1984) were obtained for the five-factor model. The likelihood-ratio chi-square test yields χ^2 (1) = .15, p = .69. Nonsignficance is considered one criterion for model acceptance discussed by Bentler and Bonett (1980). The goodness-of-fit index (Bentler & Bonett, 1980) was computed from the null model (which hypothesizes that the variables are uncorrelated) in the population, χ^2 (36) = 320.56, and the current five-factor model, χ^2 (1) = .63, as = (320.56 - .15/320.56) = .999. Thus, the model is 99% of the way to a perfect fit.

Further testing of this model included an analysis of the χ^2/df ratio, the root square residual and the Tucker-Lewis index. The χ^2/df ratio provided information on the relative efficiency of the alternative model in

accounting for the data (Marsh, Balla, & McDonald, 1988). Values of 2.0 or less were interpreted to represent an adequate fit. The present five-factor model was .15. The root mean square residual (RMSR) measured average residual correlation (Joreskog & Sorbom, 1984). Smaller values (e.g., .10 or less) reflected a better fit. The RMSR for the five-factor structure was .002. The Tucker-Lewis index (TLI) roughly scales the chi square from 0 to 1, with 0 representing the fit of the null model (Bentler & Bonett, 1980), which assumes that the variables were uncorrelated, and 1 representing the fit of a perfectly fitting model. Values less than .90 suggest that the model can be improved substantially (see Marsh et al., 1988, p. 292, for discussion), whereas values close to 1.0 indicate a better fit. This measure, when compared to the other indices, is relatively independent of sample size. The TLI in the present study was 1.10. The measure is greater than 1 because it is not standardized. Thus, the five-factor model provided an excellent representation of the data.

To interpret Table 3, we used a varimax rotation (an orthogonal solution) and considered factor loadings at or greater than .35 as meaningful. We used the common factor solution and a varimax rotation because scores on each measure had a reasonable degree of reliability and shared common variance with scores on other meas-

Table 6
Hierarchical Regression for Variables Predicting Mathematics (N=70)

	B	*SE*	*β*	*t*-ratio
Model 1				
Sem. WM	.23	.13	.08	1.73
Phon. WM	.21	.14	.13	1.50
Model 2				
Semantic Responsiveness	.62	.12	.52	5.17*
Phonological Responsiveness	.17	.13	.14	1.40
Model 3				
Verbal IQ	.81	.16	.51	4.95***
Model 4				
Semantic WM	.12	.09	.11	1.28
Phonological WM	.21	.10	.18	2.05*
Semantic Responsive	.58	.10	.49	5.75***
Phonological Responsive	.03	.10	.03	.34
Verbal IQ	.64	.13	.40	4.63***
Model 5				
Semantic WM	.16	.11	.11	1.42
Phonological WM	.15	.12	.15	1.25
Verbal IQ	.74	.17	.45	4.37***

Note. Model 1, $R^2 = .08$, $F(2,67) = 3.15$, $p < .05$; Model 2, $R^2 = .30$, $F(2,67) = 14.47$, $p < .001$; Model 3, $R^2 = .27$, $F(1,68) = 24.48$, $p < .0001$; Model 4, $R^2 = .54$, $F(5,64) = 15.17$, $p < .0001$; Model 5, $R^2 = .29$, $F(3,66) = 9.18$, $p < .0001$.
* $p < .05$, ** $p < .01$, *** $p < .001$.

ures. We were also interested in assessing the potential independent contribution of each dimension in explaining the covariation of individual differences on the measures and, therefore, used the orthogonal rotation to retain the independence of the dimensions.

The factor structure is shown in Table 4. Factor 1 showed high loadings for all semantic (digit/sentence) WM span measures across the initial, gain, and maintenance conditions. Factor 2 showed high loadings for phonological span measures (rhyming task) across the initial, gain, and maintenance condition. Thus, common variance across testing conditions emerged related to the type of WM measure used. Factor 3 showed high loadings for gain scores and probes scores for the digit/sentence (semantic) task. This factor was interpreted as reflecting reponsiveness to probes on the semantic (digit/sentence) task. In contrast, Factor 4 reflected responsiveness to probes for the rhyming (phonological) WM tasks. Factor 5 reflected reflected a single variable related to verbal intelligence.

Of interest was whether the factor scores, especially those unique to the dynamic testing format (Factors 3 and 4), contributed any unique variance to reading beyond a general verbal IQ and the semantic (digit/sentence conditions) and phonological (rhyming task conditions) WM factors (Factor 1, 2, and 5). Standard scores from the word recognition subtest of the WRAT-R served as the criterion measure in Table 5. Hierarchical regression analyses determined the amount of variance accounted for in reading scores by the five factor scores. For each model considered, variables were entered simultaneously such that the beta values reflected unique variance (the influence of all other variables partialed out). For Model 1, the semantic and phonological WM factors were entered. As shown at the bottom of Table 5, this variable accounted for approximately 16% of the variance in word recognition. The phonological factor (rhyming task conditions) contributed no significant variance in this sample.

Model 2 entered the dynamic testing factors (Factors 3 and 4) into the equation, and the total model accounted for 10% of the variance. The entry of the phonologically responsive factor contributed no significant variance. Model 3 entered the unique factor related to verbal IQ separate from the other factors. This model contributed approximately 28% of the variance to word recognition. Model 4 entered all the factor scores into the model. This accounted for approximately 44% of the variance in word recognition. As shown, this model was a substantial improvement over Model 1 ($\Delta R^2 = 28\%$) , Model 2 ($\Delta R^2 = 34\%$) and Model 3 ($\Delta R^2 = 16\%$). The results showed that only measures of semantic WM, verbal IQ, and semantic responsiveness provided unique variance in the prediction of word recognition. The final model (Model 5) removed the unique factor related to dynamic assessment (Factors 3 and 4), which reduced the variance from 44% to 38%. These results also show that dynamic testing measures (Factor 3), in addition to WM and verbal IQ, improved the prediction of word recognition. The unique contribution of the dynamic assessment factor to word recognition was 6%.

We also considered mathematics performance. Using the same models as in the reading predictions, Model 1, 2, and 3 accounted for 8%, 30%, and 27% of mathematics performance on the WRAT-R, respectively. The complete model (Model 4) indicated that the five factors accounted from approximately 54% of the variance in mathematics performance. As illustrated, this model was an improvement over Model 1 ($\Delta R^2 = 46\%$), Model 2 ($\Delta R^2 = 14\%$), and Model 3 ($\Delta R^2 = 26\%$). The results showed that only measures of phonological WM, verbal IQ, and semantic responsiveness provided unique variance in the prediction of mathematics. Thus, dynamic testing measures (Factor 3), in addition to phonological WM and verbal IQ, contributed unique variance to mathematics performance. The final model removed the unique factor related to dynamic assessment (Factors 3 and 4), resulting in a reduction of the predicted variance from 54% in the complete model (Model 4) to 29% (Model 5). Thus, the unique contribution of the dynamic assessment factor was 25%.

DISCUSSION

The purpose of this study was to determine whether dynamic assessment measures facilitate accurate classification of children with RD. Specifically, the study addressed the question of whether children classified with problems in reading or both reading and math yield different performance patterns than poor readers in responding to simple feedback (probes or cues) on tasks strongly related to achievement. The findings showed that only children with both math and reading disabilities yielded consistent patterns of weak responsiveness during dynamic testing compared to the other ability groups. Also of interest was whether dynamic testing measures contributed unique variance beyond IQ in reading. Because IQ has been considered irrelevant in predicting reading in the literature, we sought to determine whether measures of WM, as well as dynamic testing procedures, could add important variance to our prediction of reading performance. The results support our hypothesis that WM performance and dynamic testing procedures enhance our predictions of reading as well as math performance. We now discuss these two important findings.

The results yield two general findings. First, ability group differences emerged in favor of skilled readers on

both the rhyming and the sentence digit tasks. This finding is consistent with the literature linking problems in WM to poor reading (e.g., Swanson & Siegel, 2000). However, performance differences between poor readers and children with RD or MD/RD are not as straightforward and must be qualified. The present results divided performance into three areas: level of performance (as reflected in span or increasing set size scores), magnitude of change (as reflected in the magnitude of effect size), and absolute changes (as reflected in the percentage of children within groups who performance improved). In terms of the level of performance, the results showed that poor readers and children with RD or MD/RD generally performed in the same low range on verbal WM tasks. Thus, the results coincide with others' findings that poor readers and children with RD are difficult to separate on cognitive measures (e.g., Siegel, 1992; see Hoskyn & Swanson, 2000, for a review). However, these findings must be qualified because differences were found between the two groups in terms of changes in performance. For example, in terms of the magnitude of change, the results showed that effect sizes between the initial and maintenance condition were weakest for poor readers and children with MD/RD relative to skilled and children with RD. In terms of absolute changes within classification groups, however, a higher percentage of children with RD or MD/RD had maintenance scores that were the same as their initial scores. Children with RD or MD/RD were more likely to return to their initial score performance after presentation of probes (feedback) had been stopped. More specifically, approximately 60% of the children with RD and 70% of children with MD/RD failed to maintain their performance (see Table 2, column 1).

The practical implication of the above findings is that approximately 40% of the RD sample and 30% of the MD/RD were incorrectly diagnosed when change scores were taken into consideration. Stated differently, 60% of the RD and 70% of the MD/RD sample were unresponsive to dynamic testing conditions. This finding coincides with results of recent classification studies attributing treatment resistance to some children with RD (e.g., Fuchs & Fuchs, 1998; Torgesen, 2000). Further, "treatment resistance" or "treatment non-response" is becoming a key indicator of accurate classification of children with learning disabilities (Fuchs & Fuchs, 1998). Although it is unclear from our findings whether "treatment resisters" or "non-responders" under dynamic testing conditions match those who have difficulties over an extended period of intervention time (e.g., two years of intervention), the results clearly show that WM deficits related to the verbal system are less changeable for children with RD than for poor readers and skilled readers. Thus, it may be possible in the early stages of assessment to identify children at risk for RD based on their responsiveness to simple feedback.

Second, the results support the hypothesis that dynamic assessment adds significant variance in predicting reading as well as mathematics performance beyond verbal IQ. The hierarchical regression showed that dynamic assessment factor scores contributed 6% of the variance to reading and 25% to mathematics. That is, the results suggested that factors related to verbal probing and verbal gain scores for the digit/sentence task contributed unique variance to predicting reading and mathematics, beyond what is attributed to factor scores that included verbal IQ. Thus, the results support the notion that a "testing-the-limits" procedure (e.g., Carlson & Wiedl, 1979; Swanson, 1992) appears to be tapping different mental constructs than static assessment procedures (e.g., WISC-III). Further, these constructs appear to be independent of information gleaned from a traditional intelligence measure.

The findings of the present study raise a number of interesting questions. For example, why did factor scores related to verbal IQ and probing predict reading? We assume that the results reflect distinct testing conditions. Static assessment reflects two states: unaided success and failure, whereas dynamic assessment reflects some "in between" state. In the first condition, the child either answers the question correctly, without prompts or cues from the examiner, or is considered to fail the item. In the second condition, the child may be somewhere in between these two states: unable to perform the task independently but able to succeed with minimal assistance. For example, two children can earn the same low score on the S-CPT. With minimal intervention, however, one child experiences significant growth in performance, whereas the other shows little improvement. Although the two children received the same score initially, a different degree of future success may be predicted for them. This interpretation of the findings must be viewed as tentative, however, because the poor performance on either verbal IQ, reading or mathematics measures might indicate difficulty in responding to items, understanding task instructions, or a host of other factors. Further, the achievement measures reflect basic word recognition or math skills that may share a common memory construct with WM (see Swanson & Siegel, 2001, for a review). In addition, because poor readers with low verbal IQs were included in the sample, the relationship between recall of words and recall on S-CPT items may have been heightened.

The results raise an important question as to why dynamic assessment procedures are able to separate out poor readers from children operationally defined as RD or MD/RD, given that performance on the majority of

WM measures was statistically comparable for the low-achieving groups. To answer this question, let's consider the goals of dynamic assessment. Embretson (1987; also see Grigorenko & Sternberg, 1998; Swanson & Lussier, 2000; for a review) described three goals of dynamic assessment "(1) improving ability estimates, (b) assessing new constructs, and (c) improving true ability" (p. 167). These assumptions may be plausible in evaluating the sensitivity of dynamic assessment measures in predicting performance in children with MD/RD from those who are poor readers. Given these goals, we consider three explanations for the findings.

The first explanation, the one we prefer, is that the dynamic assessment measures simply provide an additional indicator of ability group differences that is not captured on static measures. An alternative explanation is that the performance differences are an artifact of spreading out the scores. It is unlikely, however, that the ability group differences were simply an artifact of inducing variance. For example, a significant number of skilled readers improved on gain conditions, whereas some children with RD did not. Therefore, instead of inducing variance or spreading out the scores, the measures were sensitive indicators of processing potential in some students and not in others.

The second explanation was that dynamic assessment measures tap new abilities: modified performance. A consideration of effect size sheds some light on whether performance was influenced by the feedback instructions provided in the gain condition. The effect size for raw scores were at least 1/2 standard deviation for gain conditions for all four ability groups. These findings suggest that responsiveness to probes was not simply an artifact of "reading ability," suggesting that some "temporal" modifications in processing performance occurred for skilled readers. We emphasize "temporal" because the readminstration of the WM tasks under maintenance conditions was more detrimental in some groups (children with MD/ RD) than others.

Finally, dynamic assessment influenced children's information-processing ability. That is, dynamic testing procedures were expected to produce changes in ability group classification because it is assumed that many psychological entities are not static (e.g., Carlson & Wiedl, 1979). The results clearly support the notion that changes in processing ability occurred across some ability groups. For example, some generally "inefficient" information processors (such as poor readers) were influenced by procedures that facilitate access to previously stored information.

In summary, the results of this study support the validity of using dynamic assessment measures to facilitate correct classification of children with RD. However, the results should be considered as preliminary because of the small sample size. Although effect sizes showing differences were adequate, not all comparisons were statistically significant. Further, other measures of processing must be developed to capture the subtle processing differences between ability groups. Although WM is an important construct related to achievement, other approaches to dynamic assessment that directly focus on academic material should be considered (e.g., Campione & Brown, 1997). However, the study demonstrates the applicability of the dynamic assessment to the measurement of learning potential and provides further evidence regarding the relationship between performance on information processing tasks and the classification of RD.

REFERENCES

Baddeley, A. D. (1986). *Working memory*. Oxford, England: Clarendon Press/Oxford University Press.

Baddeley, A. D., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105,* 158-173.

Baddeley, A. D., & Logie, R. H. (1999). The multiple-component model. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28-61). New York: Cambridge University Press.

Bateman, B. (1992). Learning disabilities: The changing landscape. *Journal of Learning Disabilities, 25,* 29-37.

Bentler, P. M., & Bonett, D. (1980). Significance test and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

Brown, A. L., & Ferrara, R. A. (1999). Diagnosing zones of proximal development. In P. Lloyd (Ed.), *L. Vygotsky: Critical assessments: the zones of proximal development* (pp. 225-256). New York: Routledge.

Brown, A. L., & French, L. A. (1979). The zone of potential development: Implications for intelligence testing in the year 2000. *Intelligence, 3,* 255-273.

Bull, R., Johnston, R. S., & Roy, J. A. (1999). Exploring the roles of the visual-spatial sketch pad and central executive in children's arithmetical skills: Views from cognition and developmental neuropsychology. *Developmental Neuropsychology, 15,* 421-442.

Campione, J. C. (1989). Assisted assessment: A taxonomy of approaches and an outline of strengths and weaknesses. *Journal of Learning Disabilities, 22,* 151-165.

Campione, J. C., & Brown, A. L. (1987). Linking dynamic testing with school achievement. In C. S. Lidz (Ed.), *Dynamic testing* (pp. 82-115). New York: Guilford Press.

Carlson, J. S., & Wiedl, K. H. (1979). Toward a differential testing approach: Testing the limits employing the Raven matrices. *Intelligence, 3,* 323-344.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19,* 450-466.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review, 3,* 442-433.

Day, J. D., Engelhardt, J. L., Maxwell, S. E., & Bolig, E. E. (1997). Comparison of static and dynamic assessment procedures and

their relation to independent performance. *Journal of Educational Psychology, 89*(2), 358-268.

De Beni, R., Palladino, P., Pazzaglia, F., & Cornoldi, C. (1998). Increases in intrusion errors and working memory deficit of poor comprehenders. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 51,* 305-320.

Elliott, J., & Lauchlan, F. (1997). Assessing potential – the search for the philosopher's stone? *Educational and Child Psychology, 14,* 6-16.

Embretson, S. E. (1987). Toward development of a psychometric approach. In C. Lidz (Ed.), *Dynamic assessment: Foundations and fundamentals* (pp. 141-172). New York: Guilford.

Embretson, S. E. (1992). Measuring and validating cognitive modifiability as an ability: A study in the spatial domain. *Journal of Educational Measurement, 29,* 25-50.

Engle, R. W., Kane, M. J., & Tuholski, S. (1999a). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 102-134). Cambridge, UK: Cambridge University Press.

Federal Register. (1977). *The rules of a regulatory for implementing P.L. 94-142*. Washington, DC: U.S. Government Printing Office.

Fletcher, J. M., Francis, D. J., Rourke, B. P., Shaywitz, S. E., & Shaywitz, B. A. (1992). The validity of discrepancy-based definitions of reading disabilities. *Journal of Learning Disabilities, 25,* 555-561.

Fletcher, J. M., Shaywitz, S. E., Shankweiler, D. P., Katz, L., Liberman, I.Y., Stuebing, K. K., Francis, D. J., Fowler, B., & Shaywitz, B. A. (1994). Cognitive profiles of reading disability: Comparisons of discrepancy and low achievement definitions. *Journal of Educational Psychology, 86,* 6-23

Fletcher, J. M., Lyon, G. R., Barnes, M., Stuebing, K. K., Francis, D. J., Olson, R. K., Shaywitz, S. E., & Shaywitz, B. A. (2002). Classification of learning disabilities: An evidenced-based evaluation. In R. Bradley, L. Danielson, & D. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 185-239). Mahwah, NJ: Erlbaum.

Fuchs, L. S., & Fuchs, D. (1998). Treatment validity: A unifying concept for reconceptualizing identification of learning disabilities. *Learning Disabilities Research & Practice, 13,* 204-219.

Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin, 124*(1), 75-111.

Hoskyn, M., & Swanson, H. L. (2000). Cognitive processing of low achievers and children with reading disabilities: A selective meta-analytic review of the published literature. *The School Psychology Review, 29,* 102-119.

Jastak, S., & Wilkinson, G. S. (1984). *Manual: The Wide Range Achievement Tests-Revised.* Wilmington, DE: Jastak Associates.

Jitendra, A. K., & Kameenui, E. J. (1993). Dynamic testing as a compensatory testing approach: A description and analysis. *RASE: Remedial and Special Education, l4,* 6-18.

Joreskog, K. G., & Sorbom, D. (1984). *LISREL VI: Analysis of linear structural relationships by the method of maximum likelihood.* Mooresville, IN: Scientific Software Inc.

Kline, R. B. (1998). *Principles and practice of structural equation modelling*. New York: Guildford.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity? *Intelligence, 14,* 389-433.

McCutchen, D. (2000). Knowledge, processing and working memory: Implication for a theory of writing. *Educational Psychologist, 35,* 13-23.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391-410.

Murphy, K. R., & Myors, B. (1998). *Statistical power analysis*. Mahwah, NJ: Erlbaum.

Nunnally, J., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw Hill.

Shaywitz, S. E., Fletcher, J. M., Holahan, J. M., Shneider, A. E., Marchione, K. E., Stuebing, K. K., Francis, D., & Shaywitz, B. A. (1999). Persistence of dyslexia: The Connecticut longitudinal study at adolescence. *Pediatrics, 104,* 1351-1359.

Shepherd, L. A., Smith, M. L., & Vojir, C. P. (1983). Characteristics of pupils identified as learning disabled. *American Educational Research Journal, 20,* 309-331.

Siegel, L. S. (1989). IQ is irrelevant to the definition of learning disabilities. *Journal of Learning Disabilities, 22,* 469-478.

Siegel, L. S. (1992). An evaluation of the discrepancy definition of dyslexia. *Journal of Learning Disabilities, 25,* 618-629.

Siegel, L. S., & Ryan, E. B. (1989). The development of working memory in normally achieving and subtypes of learning disabled. *Child Development, 60,* 973-980.

Stanovich, K., & Siegel, L.S. (1994). Phenotypic performance profile of children with reading disabilities: A regression-based test of the phonological-core variable-difference model. *Journal of Educational Psychology, 86,* 24-53.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Erlbaum.

Stuebing, K. K., Fletcher, J. M., LeDoux, J. M., Lyon, G. R., Shaywitz, S. E., & Shaywitz, B. A. (2002). Validity of IQ-discrepancy classifications of reading disabilities: A meta-analysis. *American Educational Research Journal, 39,* 469-518.

Swanson, H. L. (1992). Generality and modifiability of working memory among skilled and less skilled readers. *Journal of Educational Psychology,* 84, 473-488.

Swanson, H. L. (1993). Working memory in learning disability subgroups. *Journal of Experimental Child Psychology, 56,* 87-114.

Swanson, H. L. (1995a). *Swanson-Cognitive Processing Test (S-CPT).* Austin, TX: Pro-Ed.

Swanson, H. L. (1995b). Using the Cognitive Processing Test to assess ability: Development of a dynamic measure. *School Psychology Review, 24*, 672-693.

Swanson, H. L. (1996). Individual and age-related differences in children's working memory. *Memory & Cognition, 24,* 70-82.

Swanson, H. L. (1999). Reading comprehension and working memory in skilled readers: Is the phonological loop more important than the executive system? *Journal of Experimental Child Psychology, 72,* 1-31.

Swanson, H. L. (2003). Age-related differences in learning disabled and skilled readers' working memory. *Journal of Experimental Child Psychology, 85,* 1-31.

Swanson, H. L., & Hoskyn, M. (1998). Experimental intervention research on students with learning disabilities: A meta-analysis of treatment outcomes. *Review of Educational Research, 68,* 277-321.

Swanson, H. L., & Lussier, C. (2001). A selective synthesis of the experimental literature on dynamic assessment. *Review of Educational Research, 71,* 321-363.

Swanson, H. L., & Siegel, L. (2001). Learning disabilities as a working memory deficit. *Issues in Education: Contributions from Educational Psychology, 7,* 1-48.

Torgesen, J. K. (2002). Empirical and theoretical support for direct diagnosis of learning disabilities by assessment of intrinsic processing weaknesses. In R. Bradley, L. Danielson, & D. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 565-603). Mahwah, NJ: Erlbaum.

Torgesen, J. K. (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research and Practice, 15,* 55-64.

Vygotsky, L. S. (1978). Interaction between learning and development. In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.), *Mind in society: The development of higher psychological processes* (pp. 79-91). Cambridge, MA: Harvard University Press (original work published 1935).

Wechsler, D. (1991). *Manual: Wechsler Intelligence Scale for Children-Third Edition.* New York: Harcourt Brace Jovanovich, Inc.

Ysseldyke, J. E., Algozzine, B., Shinn, M. R., & McGrue, M. (1982). Similarities and differences between low achievers and students classified learning disabled. *Journal of Special Education, 16,* 73-85.

FOOTNOTES

1. Everyday examples of WM tasks would thus include holding a person's address in mind while listening to instructions about how to get there, or listening to the sequence of events in a story while trying to understand what the story means. This is to be contrasted with everyday examples of STM tasks that include recalling a series of digits in order, such as a telephone number, immediately after their presentation. Although there is controversy concerning the nature of STM and WM tasks, there is some agreement that a transformation or active monitoring (i.e., focusing on relevant information when competing information is present) is required on WM tasks (e.g., Baddeley & Logie, 1999). This monitoring draws resources from long-term memory. For the sake of parsimony, the present study views WM tasks as those that require some inference, transformation, and monitoring of relevant and irrelevant information. In contrast, STM tasks require the storage of information with minimal "ongoing" processing requirements that vary from initial encoding.
2. The purpose of this study was to determine whether DA contributes unique variance in predicting reading. However, it is also possible that DA procedures contribute to current models of reading disabilities as they are related to WM. WM involves two activities: processing and storage (e.g., Daneman & Carpenter, 1980). It has recently been argued that children with reading and math disabilities demonstrate WM deficits related to the storage of verbal information (e.g., Swanson, 2003; Swanson & Siegel, 2001). Although their processing efficiency can be improved upon, constraints in storage limit any substantial changes in WM performance. On the assumption that the RD reflects constraints in storage, DA procedures may assist in separating the children with difficulties in storage from children who suffer processing inefficiencies. Because DA improves processing efficiency (i.e., improved retrieval of stored information via feedback and prompts), one could argue from the extant literature that children with RD demonstrate less change than poor readers. As will be shown in this study, children with RD responded favorable to DA procedures; however, they were less likely than the poor readers to "hold" or "maintain" those positive changes over time.
3. A recent meta-analysis (Hoskyn & Swanson, 2000) found that the WRAT-R was the most common instrument used to compare RD and poor readers on measures of reading.
4. Criteria for selecting the poor reading sample were also based on an attempt to match reading scores with verbal IQ scores as well to statistically match the poor reading sample to RD children on reading scores. Finally, the cut-off score for the poor readers was based on a frequently cited study by Ysseldyke, Algozzine, Shinn, and McGue (1982), which compared LD and low-achieving children on several standardized achievement measures.

Requests for reprints should be addressed to: H. Lee Swanson, Educational Psychology/Special Education, University of California, Riverside, CA 92521; lee.swanson@ucr.edu.